Supplementary material

POLYPHEMUS: R package for comparative analysis of RNA Polymerase II ChIP-seq profiles by non-linear normalization

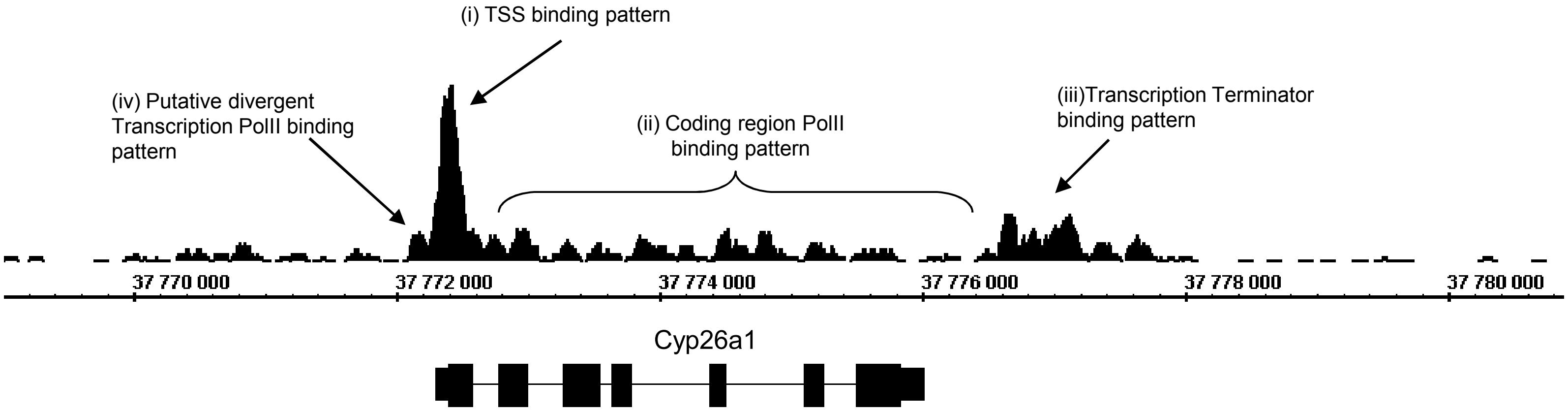
Marco A. Mendoza-Parra[1, *], Martial Sankar[1,$], Mannu Walia and Hinrich Gronemeyer*
Department of Cancer Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)/CNRS/INSERM/Université de Strasbourg, BP 10142, 67404 Illkirch Cedex, France

[$]présent address : Department of Plant Molecular Biology, University of Lausanne, Biophore Building, CH-1015 Lausanne, Switzerland.
[1]These authors have contributed equally

*Correspondance:

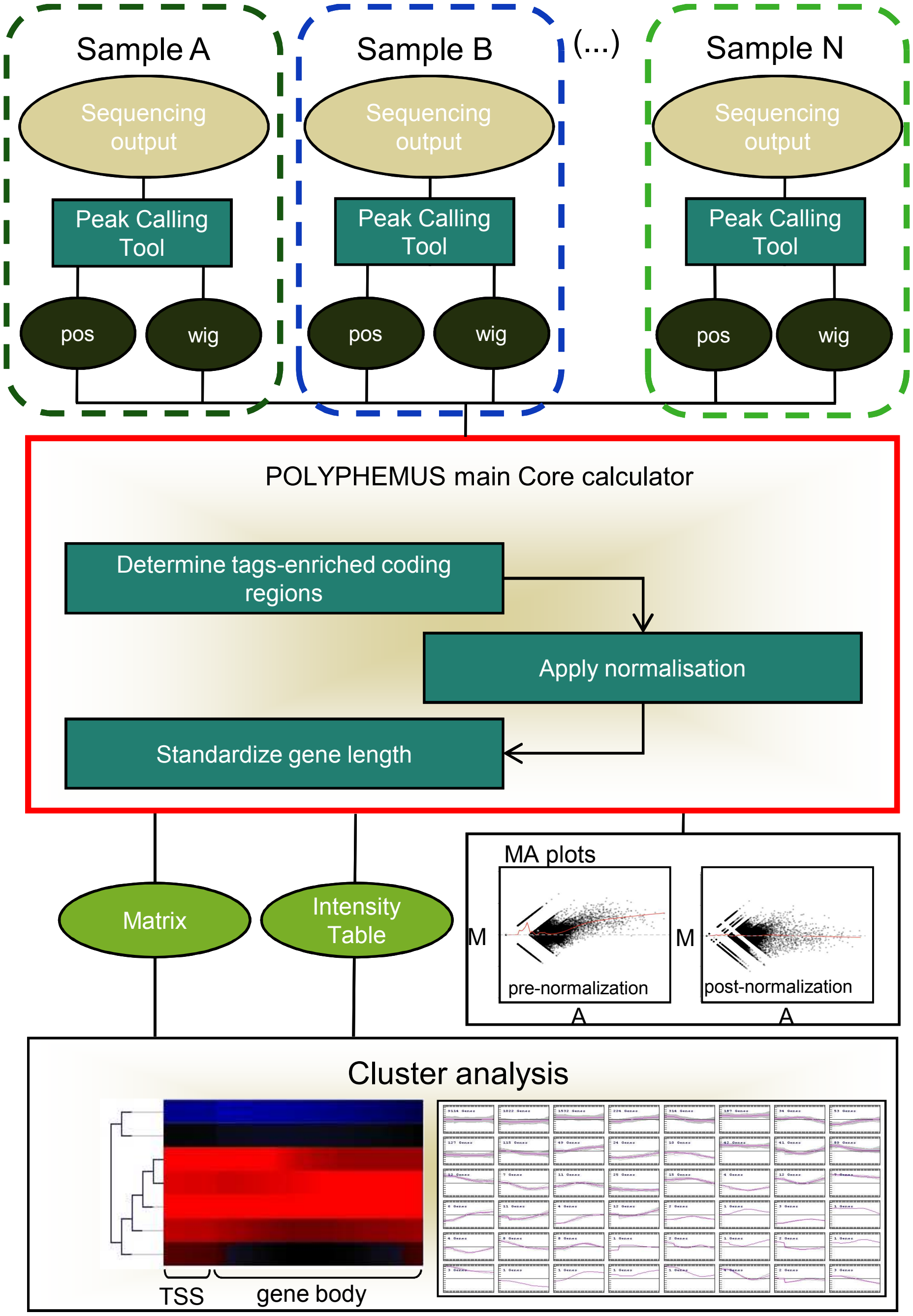
Marco A. Mendoza-Parra
Email: marco@igbmc.fr

Hinrich Gronemeyer
E-mail: hg@igbmc.fr
Phone: +(33) 3 88 65 34 73
Fax: +(33) 3 88 65 34 37



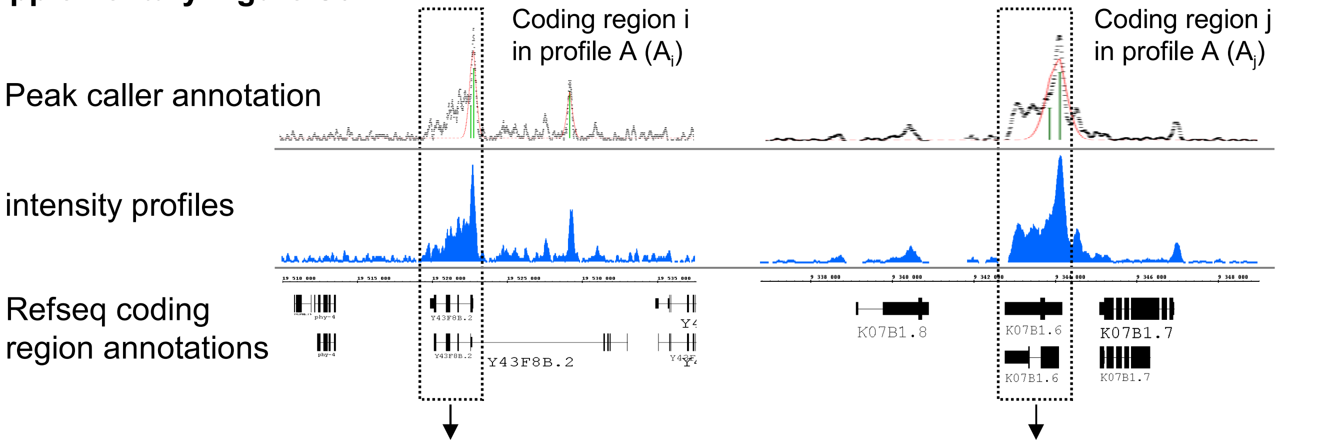
Supplementary Figure S1 – Typical RNA Polymerase II pattern observed in ChIP-sequencing profiles

In contrast to Transcription Factors, RNA PolII presents a complex chromatin interaction behavior as reflected on its ChIP-seq profile: (i) An strong signal intensity pattern localized at the TSS; (ii) an RNA PolII signal enrichment over the coding region that my be directly proportional to the transcriptional activity; (iii) In certain cases is possible to observe an stronger signal enrichment at the end of the coding region, which can be associated to the transcription terminator region; (iv) A rather small defined peak located upstream of the TSS which may reflect the presence of a divergent transcriptional pattern.



Supplementary Figure S2 – Principle of RNA Polymerase II ChIP-seq profiles comparison
POLYPHEMUS integrates read-count signal intensity profiles (“wig” for wiggle format file) with PolII binding site annotations (“pos” for significant peak position) to identify PolII-associated coding regions. This process, which can be performed with multiple ChIP-seq data sets (“Samples A to N”), generates the input for the main POLYPHEMUS calculator, which (i) extracts the coding regions of interest (ii) performs a non-linear normalization of all compared datasets and (iii) standardizes gene lengths for intra-coding region comparisons. Subsequently, MA plots are generated to validate the normalization procedure; an Intensity Table is generated to recall the transformations applied to the coding regions of interest and a Matrix Table associate them to relative signal intensity ratios for the regions of interest. These tables are used for further classification analysis performed with tools like MultiExperiment Viewer, MeV(Saeed et al. (2003), Biotechniques 34, 374-8);(Saeed et al. (2006), Methods Enzymol. 411, 134-93), such as the cluster analysis illustrated in the figure.

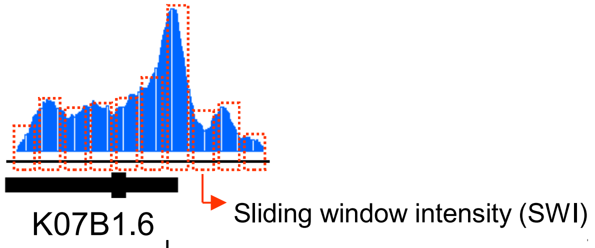
Supplementary Figure S3.

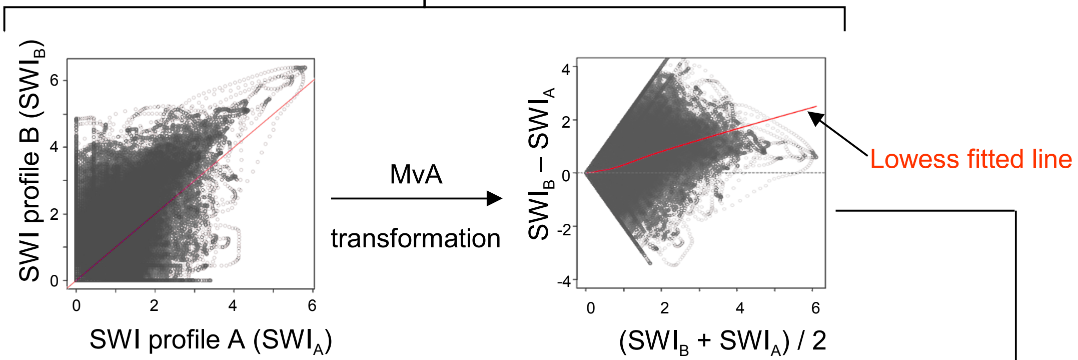


1. PolII-enriched predicted sites overlapping with transcription start site (TSS) annotated regions are selected.

2. Signal intensity information retrieved over the selected coding regions are extracted.

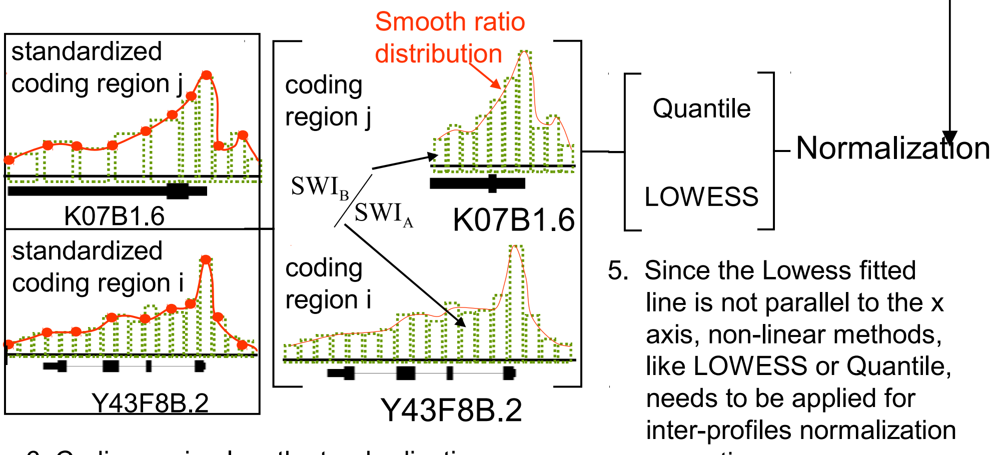
3. Signal intensity information is collected in 250bp sliding windows. The median intensity per sliding window is defined as SWI





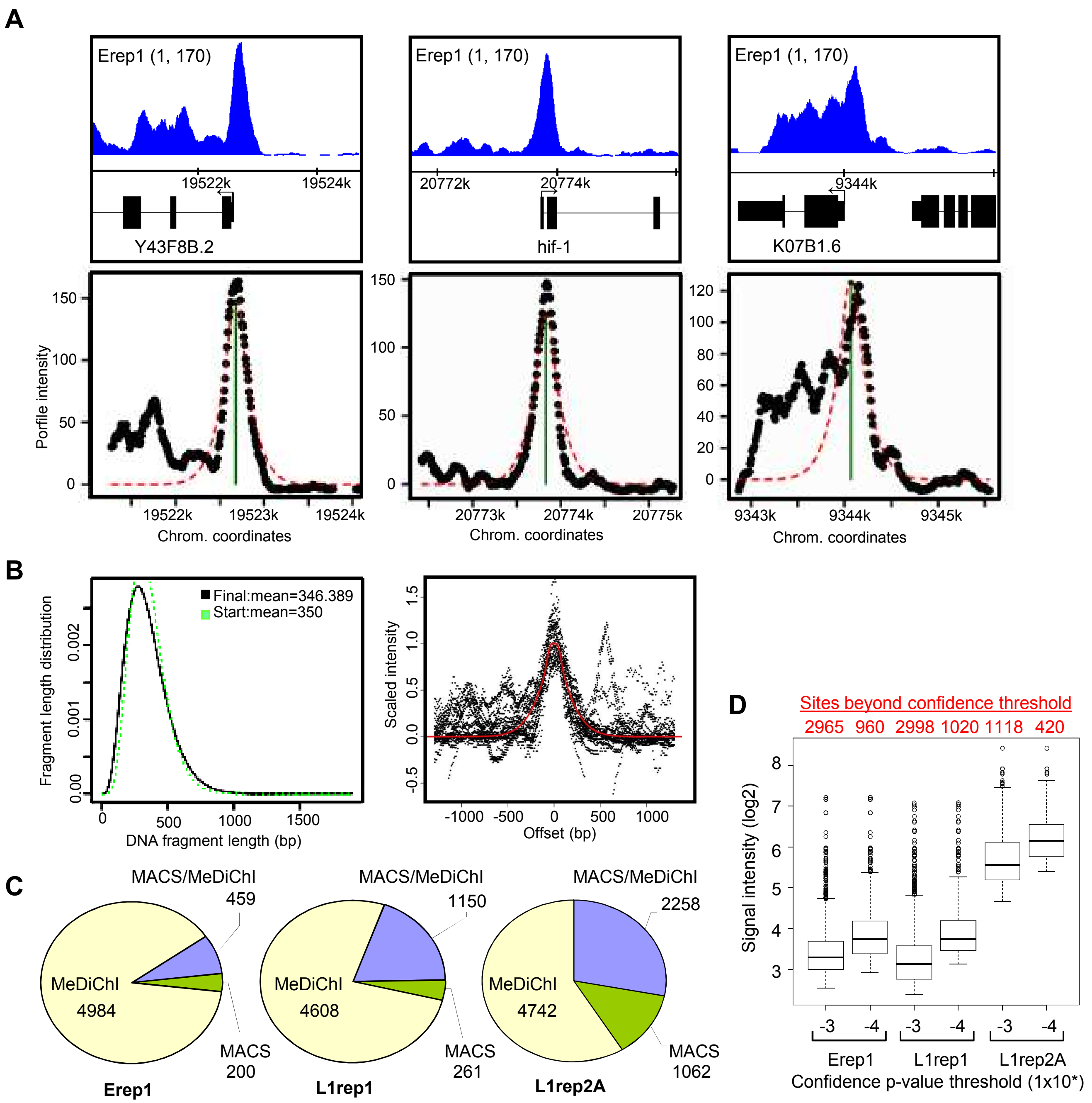
4. Given two profiles to be compared, A and B, their SWI are displayed in a MvA plot to assess whether their global behaviour requires to be normalized. A Lowess fitted line, which represent the mean bahaviour of the population is used to illustrate the SWI bias between compared profiles.

7.POLYPHEMUS normalized profiles are expressed in a Differential PolII enrichment manner; e.i. as a ratio between the SSWI associated to the compared coding regions. This information can be further used for coding regions classifications based on their relative PolII enrichment between compared profiles.



5. Since the Lowess fitted line is not parallel to the x axis, non-linear methods, like LOWESS or Quantile, needs to be applied for inter-profiles normalization corrections

6. Coding region length standardization



Supplementary Figure S4- Peak detection strategy for RNA-PolII profiles using MeDiChI.

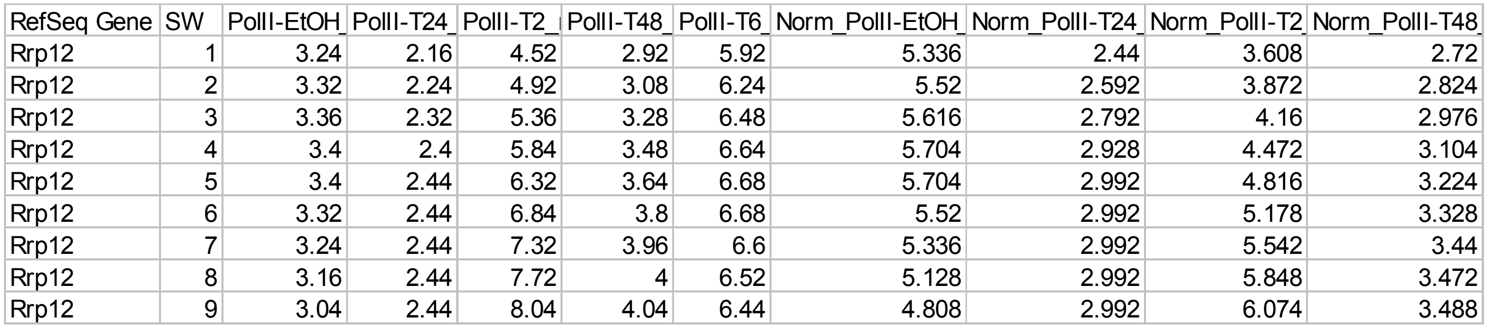
A. The deconvolution approach performed by MeDiChI (Reiss, Facciotti et al. (2008), Bioinformatics 24, 396-403) allows the proper annotation of Transcription Start Sites (TSS) in RNA-PolII ChIP-seq profiles even in cases where the density of the signal on the neighboring coding regions makes its read-out difficult. Three examples of coding regions differently occupied by RNA PolII in the embryo-stage ChIP-seq samples (top panel) are shown. The Fitted model distribution by MeDiChI (bottom panel, red line) illustrates the identification of the TSS sites in these three examples. **B.** The predicted DNA fragmentation size based on the modelled distribution associated to the ChIP-seq samples, as well as the Kernel modelled distribution used for peak detection (left and right panel respectively) are also represented. **C.** Comparison between MACS (Zhang, Y. et al. (2008), Genome Biol. 9, R137), a widely used ChIP-seq peak caller and MeDiChI in the context of their performance for annotating known TSS PolII enriched regions. For the conditions used (confidence p-value threshold for MACS and MeDiChI are 1x10^{-5} and 1x10^{-2} respectively) MeDiChI identifies more TSS occupied sites than MACS, but in addition the population sizes predicted for different samples (Erep1; L1rep1; L1rep2A) are quite similar (relative standard deviation RSD =13.5%) in contrast to that predicted by MACS (RSD= 76%). **D.** PolII-enrichment signal intensities for TSS sites exclusively predicted by MeDiChI at different confidence p-value threshold conditions. Note that MACS and MeDiChI use different approaches for estimating the confidence indicators; for this reason the comparative analysis cannot be performed at identical p-values. MACS p-value range: 1x10^{-5} to 8x10^{-106}; MeDiChI p-value range: 1x10^{-1} to 4x10^{-5}.

POLYPHEMUS intensity table

Compared samples prior norm (SWI)

Compared samples after normalisation (SWI)

Sliding windows per enriched coding regions



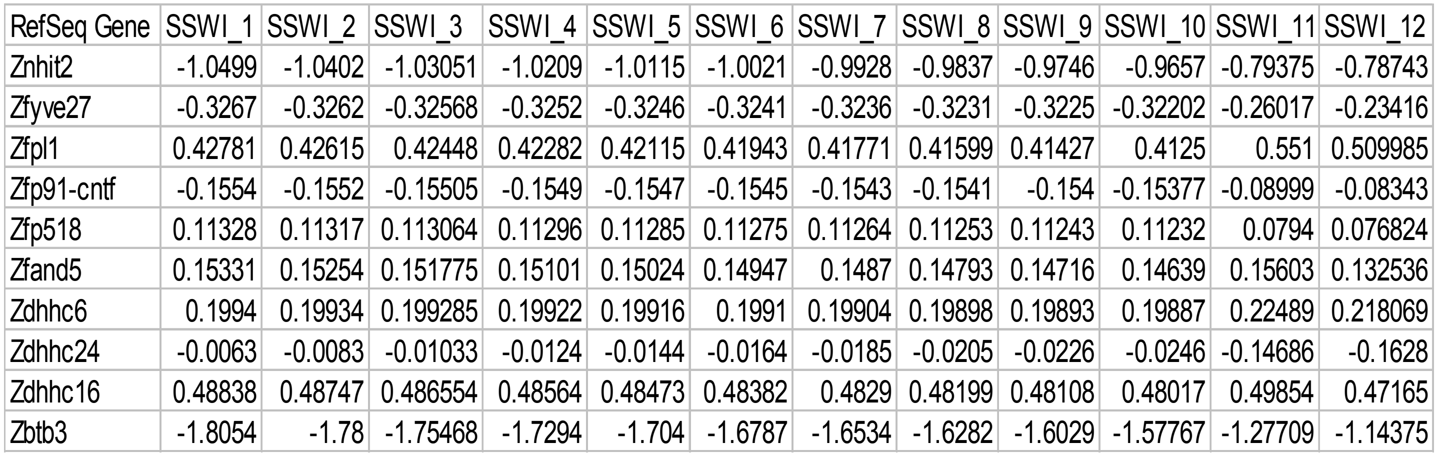
RefSeq Gene	SW	PolII-EtOH_	PolII-T24_	PolII-T2_	PolII-T48_	PolII-T6_	Norm_PolII-EtOH_	Norm_PolII-T24_	Norm_PolII-T2_	Norm_PolII-T48_
Rrp12	1	3.24	2.16	4.52	2.92	5.92	5.336	2.44	3.608	2.72
Rrp12	2	3.32	2.24	4.92	3.08	6.24	5.52	2.592	3.872	2.824
Rrp12	3	3.36	2.32	5.36	3.28	6.48	5.616	2.792	4.16	2.976
Rrp12	4	3.4	2.4	5.84	3.48	6.64	5.704	2.928	4.472	3.104
Rrp12	5	3.4	2.44	6.32	3.64	6.68	5.704	2.992	4.816	3.224
Rrp12	6	3.32	2.44	6.84	3.8	6.68	5.52	2.992	5.178	3.328
Rrp12	7	3.24	2.44	7.32	3.96	6.6	5.336	2.992	5.542	3.44
Rrp12	8	3.16	2.44	7.72	4	6.52	5.128	2.992	5.848	3.472
Rrp12	9	3.04	2.44	8.04	4.04	6.44	4.808	2.992	6.074	3.488

POLYPHEMUS normalized SSWI ratios between two compared PolII ChIP-seq profiles

Normalized SSWI ratio TSS

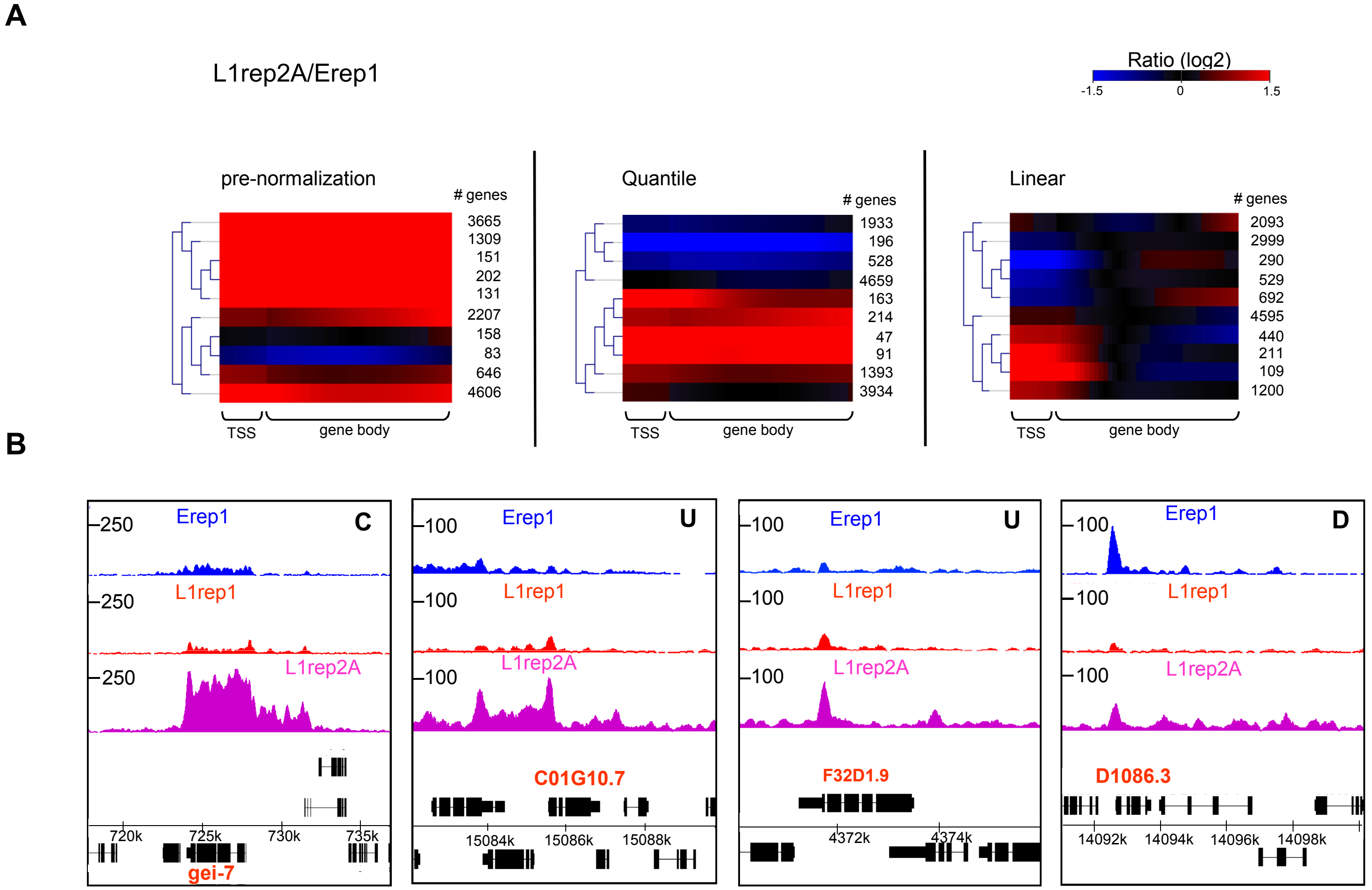
Normalized SSWI ratio gene body (SSWI11…SSWI50)

PolII enriched coding regions



RefSeq Gene	SSWI_1	SSWI_2	SSWI_3	SSWI_4	SSWI_5	SSWI_6	SSWI_7	SSWI_8	SSWI_9	SSWI_10	SSWI_11	SSWI_12
Znhit2	-1.0499	-1.0402	-1.03051	-1.0209	-1.0115	-1.0021	-0.9928	-0.9837	-0.9746	-0.9657	-0.79375	-0.78743
Zfyve27	-0.3267	-0.3262	-0.32568	-0.3252	-0.3246	-0.3241	-0.3236	-0.3231	-0.3225	-0.32202	-0.26017	-0.23416
Zfpl1	0.42781	0.42615	0.42448	0.42282	0.42115	0.41943	0.41771	0.41599	0.41427	0.4125	0.551	0.509985
Zfp91-cntf	-0.1554	-0.1552	-0.15505	-0.1549	-0.1547	-0.1545	-0.1543	-0.1541	-0.154	-0.15377	-0.08999	-0.08343
Zfp518	0.11328	0.11317	0.113064	0.11296	0.11285	0.11275	0.11264	0.11253	0.11243	0.11232	0.0794	0.076824
Zfand5	0.15331	0.15254	0.151775	0.15101	0.15024	0.14947	0.1487	0.14793	0.14716	0.14639	0.15603	0.132536
Zdhhc6	0.1994	0.19934	0.199285	0.19922	0.19916	0.1991	0.19904	0.19898	0.19893	0.19887	0.22489	0.218069
Zdhhc24	-0.0063	-0.0083	-0.01033	-0.0124	-0.0144	-0.0164	-0.0185	-0.0205	-0.0226	-0.0246	-0.14686	-0.1628
Zdhhc16	0.48838	0.48747	0.486554	0.48564	0.48473	0.48382	0.4829	0.48199	0.48108	0.48017	0.49854	0.47165
Zbtb3	-1.8054	-1.78	-1.75468	-1.7294	-1.704	-1.6787	-1.6534	-1.6282	-1.6029	-1.57767	-1.27709	-1.14375

Supplementary Figure S5. Examples of POLYPHEMUS generated outputs.



Supplementary Figure S6 – Comparison between quantile and linear normalization approaches

A. Comparison between L1rep2A (~8 million reads) and Erep1 (~2 million reads) without performing signal intensity normalization leads to the misleading interpretation of a strong upregulation transcription for the Larvae-L1 stage (> 90% of the genes present a fold change ratio higher than 1, as illustrated by the red color in the heat map of the left panel). On the other hand, the linear scaling correction shows events in which the TSS are PolII enriched (red) but the corresponding gene body appears to be depleted (blue) whereas other genes displays the opposite behavior (right panel). Only the non-linear quantile approach can properly separate RNA PolII enriched (red) from depletion (blue) events (middle panel) as it was observed when comparing L1rep1/Erep1 (see Figure 3).

B. Examples of different PolII-associated annotated genes identified after quantile normalization. The signal intensity tracks for Erep1, L1rep1 and L1rep2A are not normalized; the indicated genes (red label) were predicted as constitutive (Gei-7), up-regulated (C01G10.7; F32D1.9) or down-regulated (D1086.3) for PolII transcriptional activity. Note that the biological replicate L1rep2A cannot be easily compared with L1rep1 under the represented conditions.

Supplementary Figure S7 –Z-transformation for Normalized RNA Polymerase ChIP-seq data.

Standardized sliding window intenstity ratios (SSWI) Matrix generated by POLYPHEMUS has been transformed in a way that the distribution of the median SSWI ratios per coding region is fitted into the standard normal distribution. This is performed by the following expression:

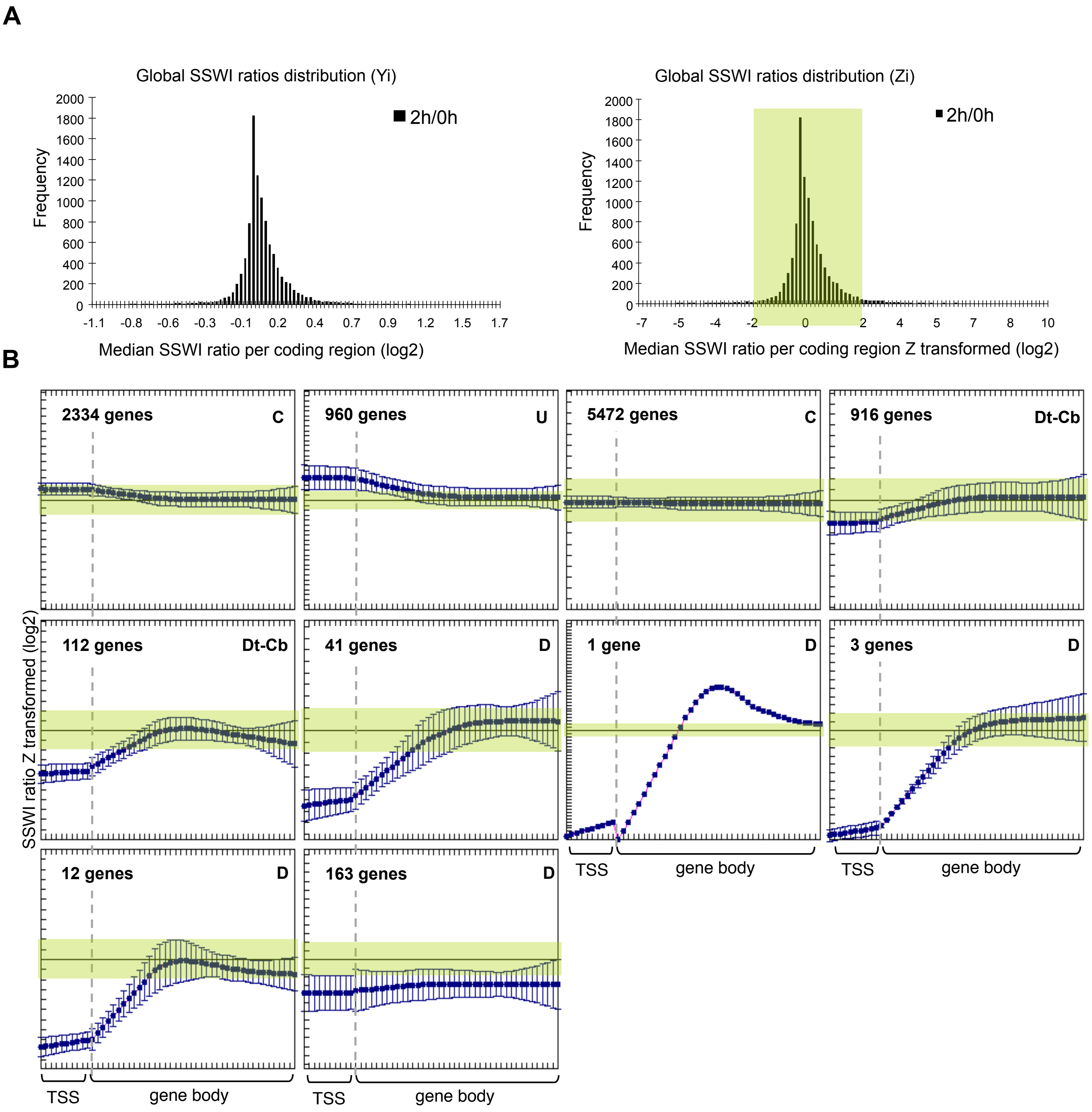
$$\begin{pmatrix} \text{Gene\_1 } \text{SSWIr\_1,1} \ldots \text{SSWIr\_1,n} \\ \\ \text{Gene\_i } \text{SSWIr\_i,1} \ldots \text{SSWIr\_i,n} \\ \\ \text{Gene\_m } \text{SSWIr\_m,1} \ldots \text{SSWIr\_m,n} \end{pmatrix} \begin{matrix} \rightarrow \text{Yi}=\overset{j=1:n}{\text{median}}(\text{SSWIr\_1,j}) \\ \\ \rightarrow \text{Yi}=\overset{j=1:n}{\text{median}}(\text{SSWIr\_i,j}) \\ \\ \rightarrow \text{Yi}=\overset{j=1:n}{\text{median}}(\text{SSWIr\_m,j}) \end{matrix} \xrightarrow[\text{Z transformation}]{\text{Zi}=(\text{Yi}-\mu)/\sigma} \begin{matrix} \text{Z1} \rightarrow \\ \\ \text{Zi} \rightarrow \\ \\ \text{Zm} \rightarrow \end{matrix} \begin{pmatrix} \text{Gene\_1 } \text{Z\_SSWIr\_1,1} \ldots \text{Z\_SSWIr\_1,n} \\ \\ \text{Gene\_i } \text{Z\_SSWIr\_i,1} \ldots\ldots \text{Z\_SSWIr\_i,n} \\ \\ \text{Gene\_m } \text{Z\_SSWIr\_m,1} \ldots \text{Z\_SSWIr\_m,n} \end{pmatrix}$$

$$\mu=\frac{\sum_{i=1}^{m}(\text{Yi})}{m} \qquad \sigma=\sqrt{\frac{\sum_{i=1}^{m}(\text{Yi}-\mu)^{2}}{m}}$$

$$\text{Z\_SSWIr\_i,j}=(\text{SSWIr\_i,j}-\mu)/\sigma$$

The following figures show examples in which this transformation was applied in order to study the temporal chromatin association of RNA polymerase monitored by ChIP-seq during F9 embryonal carcinoma cell differentiation upon 2h, 6h, 24h and 48h exposure to ATRA.

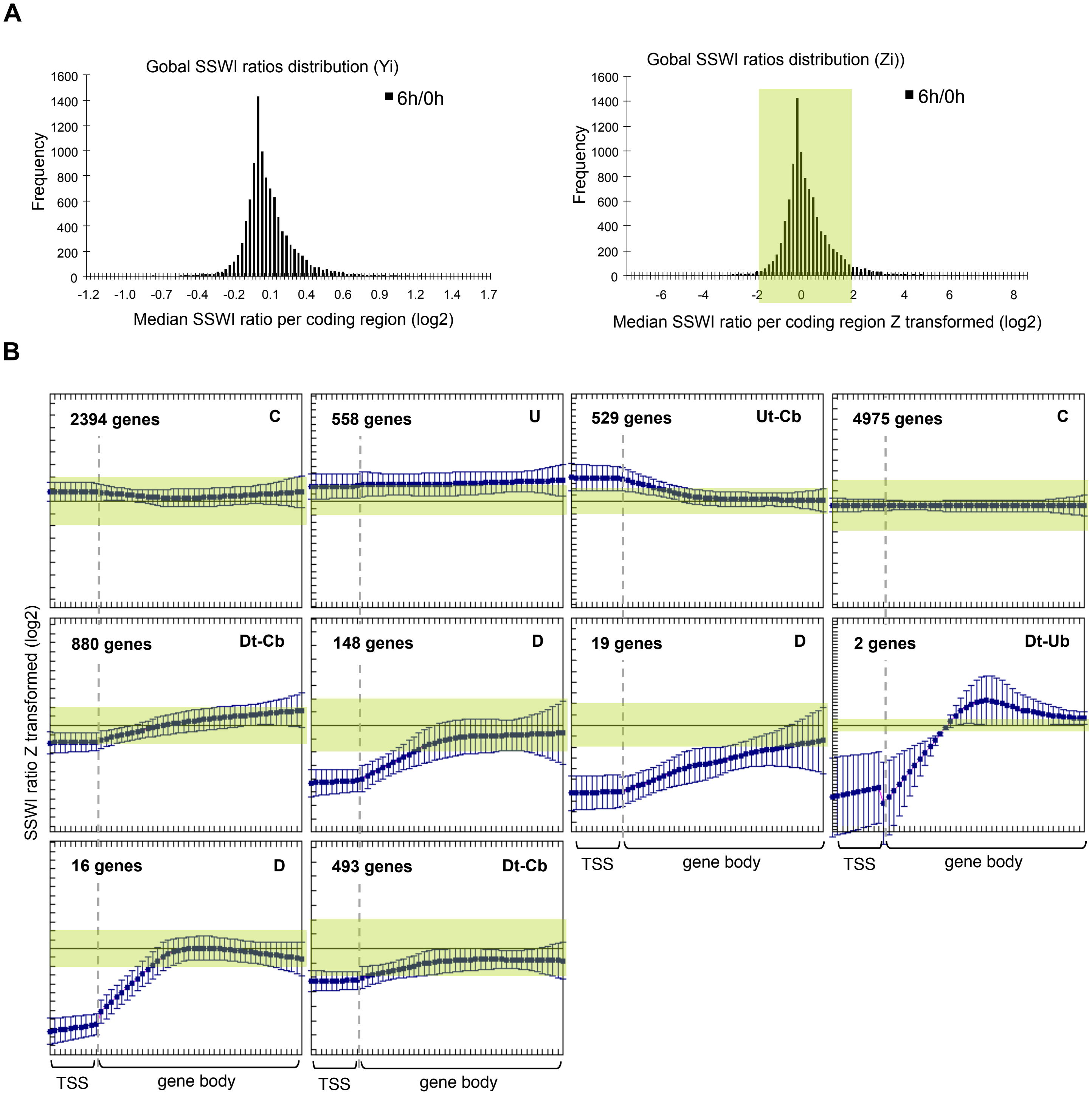
RNA Polymerase II ChIP-Seq comparison: 2h/0h ATRA treated F9 cells (QUANTILE normalized ratios)



A. The Global SSWI ratios distribution before and after Z-transformation is illustrated below for each compared profiles (left and right panel respectively). Notice that in the corresponding standard normal distribution is illustrated the significance level of 5% of probability for the detection of differential transcriptional behaviors (+/-2σ distance away from the global mean behavior; outside of the shadowed region).

B. The Z-transformed Matrix has been used for SOTA classification (Max Cycles=9; Cell variability p-value= 0.01) and the different groups are illustrated by centroid graphs (the corresponding Heatmaps are illustrated in figure 5A). Each group has been classified according to the ratio levels observed at the TSS as well as at the gene body (following the classification described in figure 3) taking in consideration a 5% confidence criteria.

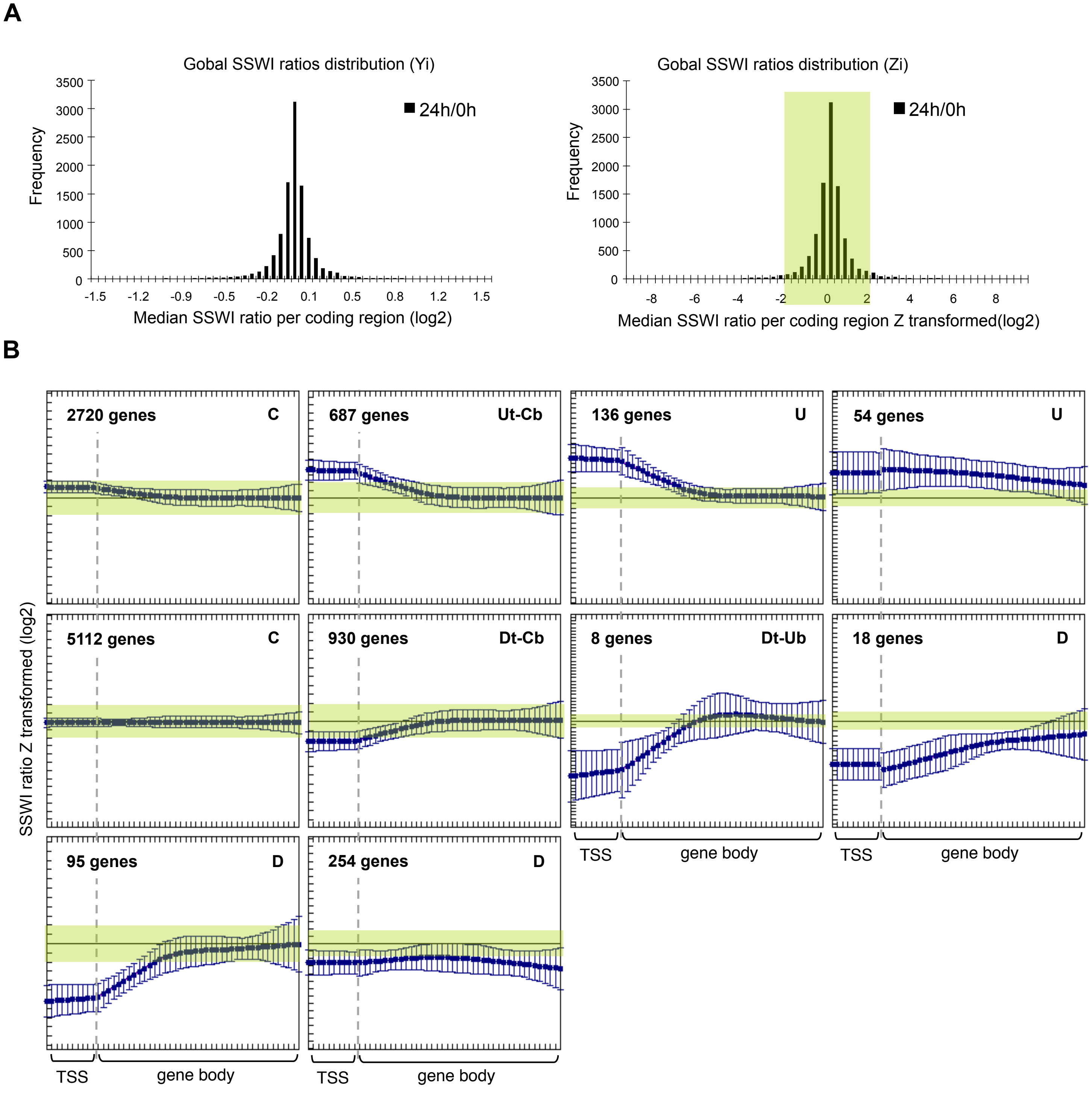
RNA Polymerase II ChIP-Seq comparison: 6h/0h ATRA treated F9 cells (QUANTILE normalized ratios)



A. The Global SSWI ratios distribution before and after Z-transformation is illustrated below for each compared profiles (left and right panel respectively). Notice that in the corresponding standard normal distribution is illustrated the significance level of 5% of probability for the detection of differential transcriptional behaviors (+/-2σ distance away from the global mean behavior; outside of the shadowed region).

B. The Z-transformed Matrix has been used for SOTA classification (Max Cycles=9; Cell variability p-value= 0.01) and the different groups are illustrated by centroid graphs (the corresponding Heatmaps are illustrated in figure 5A). Each group has been classified according to the ratio levels observed at the TSS as well as at the gene body (following the classification described in figure 3) taking in consideration a 5% confidence criteria.

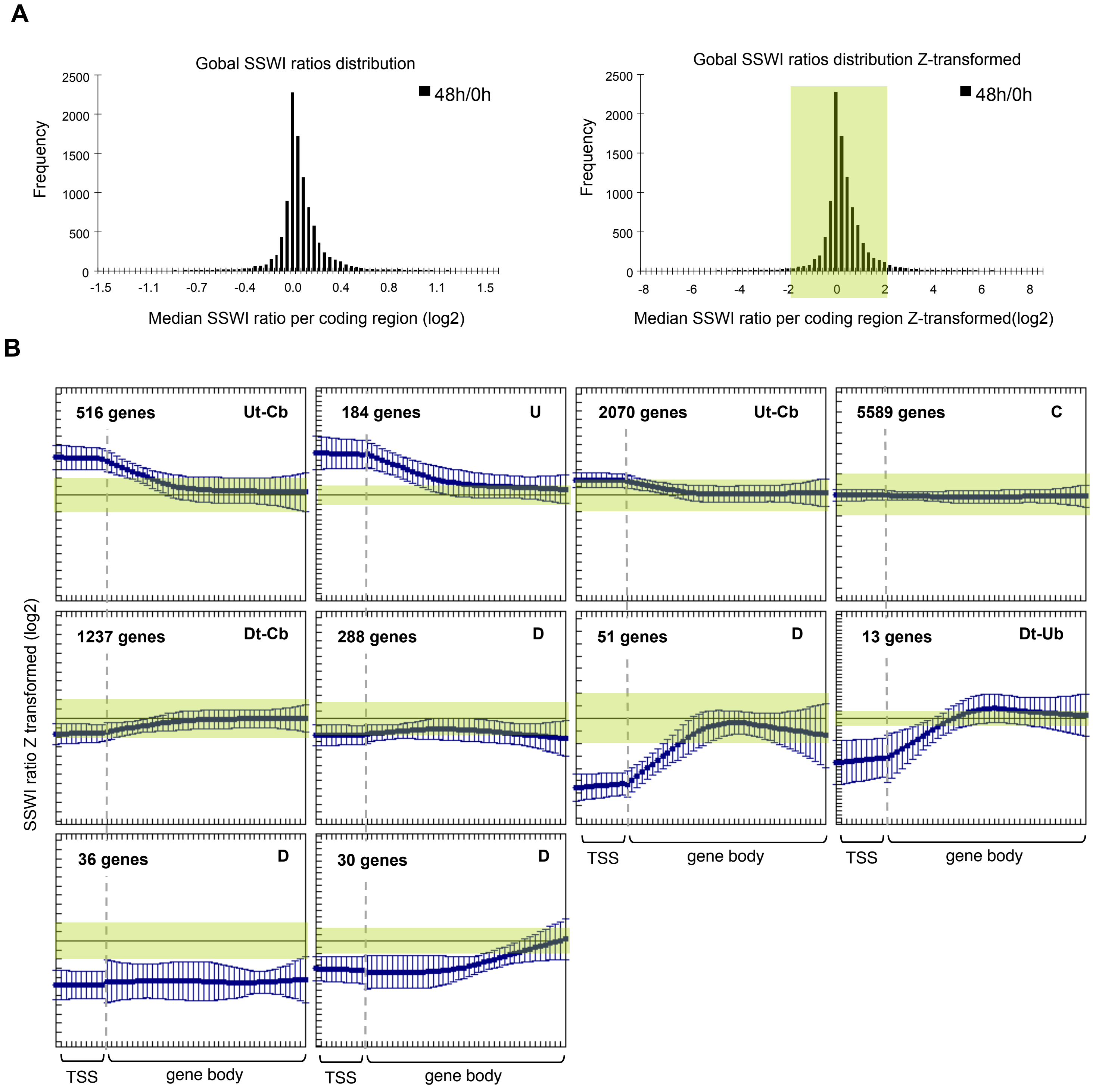
RNA Polymerase II ChIP-Seq comparison: 24h/0h ATRA treated F9 cells (QUANTILE normalized ratios)



A. The Global SSWI ratios distribution before and after Z-transformation is illustrated below for each compared profiles (left and right panel respectively). Notice that in the corresponding standard normal distribution is illustrated the significance level of 5% of probability for the detection of differential transcriptional behaviors (+/-2σ distance away from the global mean behavior; outside of the shadowed region).

B. The Z-transformed Matrix has been used for SOTA classification (Max Cycles=9; Cell variability p-value= 0.01) and the different groups are illustrated by centroid graphs (the corresponding Heatmaps are illustrated in figure 5A). Each group has been classified according to the ratio levels observed at the TSS as well as at the gene body (following the classification described in figure 3) taking in consideration a 5% confidence criteria.

RNA Polymerase II ChIP-Seq comparison: 48h/0h ATRA treated F9 cells (QUANTILE normalized ratios)



A. The Global SSWI ratios distribution before and after Z-transformation is illustrated below for each compared profiles (left and right panel respectively). Notice that in the corresponding standard normal distribution is illustrated the significance level of 5% of probability for the detection of differential transcriptional behaviors (+/-2σ distance away from the global mean behavior; outside of the shadowed region).

B. The Z-transformed Matrix has been used for SOTA classification (Max Cycles=9; Cell variability p-value= 0.01) and the different groups are illustrated by centroid graphs (the corresponding Heatmaps are illustrated in figure 5A). Each group has been classified according to the ratio levels observed at the TSS as well as at the gene body (following the classification described in figure 3) taking in consideration a 5% confidence criteria.

Supplementary_File S1

RNA Polymerase II ChIP-seq profiling during ATRA-induced F9 cell differentiation

Method:

1.5 million F9 cells were seed in 15cm plated containing Dulbecco's modified Eagle's medium supplemented with 10% foetal calf serum. All-trans retinoic acid (ATRA) was added to the plates in a final concentration of 1x10^{-6} M and this at different time points as following:

- 48 hours treatment: After seeding
- 24 hours treatment: 24 hours after seeding
- 6 hours treatment: 42 hours after seeding
- 2 hours treatment: 46 hours after seeding
- 48 hours control: after seeding, Ethanol is added instead of ATRA

After 48 hours, cells are fixed with Paraformaldehyde (Electron Microscopy Sciences) 1% final concentration during 30 minutes, washed three times with 1xPBS buffer and collected for ChIP assay.

ChIP assay is following standard procedures. Briefly, RNA PolII antibody (Santa Cruz Biotechnology; sc-9001 (H-224)) is incubated overnight with 3 million cells whole-cell extracts previously resuspended on Lysis buffer (Tris-Cl 50mM, NaCl 140mM, EDTA 1mM, Triton-X 1%, Sodium Deoxycholate 0.1%, supplemented with Protease inhibitors: Complete EDTA-free, Roche ref: 11873580001) and fragmented by sonication until an average length of 200-500bps. Next day 25ul of Protein A Sepharose beads (SIGMA, ref:P9424) previously blocked with BSA (5mg/ml) and tRNA are added and incubated during at least one hour. Then, protein A beads are washed twice with Lysis buffer, Lysis buffer 500Mm NaCl, washing buffer (EDTA 1mM, Tris-Cl 10mM, LiCl 250mM, NP40 0.5%, Sodium Deoxycholate 0.5%) and once with 1XTE (10Mm Tris-Cl pH 8.0; 1Mm EDTA pH 8.0). DNA is eluted from beads with Elution buffer (SDS 1%; TrisCl 50mM pH 8.0; EDTA 10mM) at 65°C during 15 minutes. The immunoprecipitated DNA has been quantified using Qubit (Quant-iT dsDNA HS Assay Kit; Invitrogen). 10ng of the Immunoprecipitated material is used as substrate for sequencing library preparation.

After quantifying the prepared library with the Agilent 2100 Bioanalyser, 5pmol of it were used per channel of the SOLEXA 1G Genome Analyser (Illumina). Between 10 to 15 million clusters were generated using the Illumina single read cluster generation Kit V2 at the end of a sequencing run comprising 36-cycles. The Illumina Pipeline v1.4.0 comprising image analysis (Firecrest), base calling (Bustard) and alignment (Gerald) steps was used for initial data analysis. Uniquely aligned tags with up to two mismatches relative to the mm9 mouse reference genome were kept. The total number of mapped reads is listed on figure 4.